1

Twelve Steps for Effective Test Development

Steven M. Downing
University of Illinois at Chicago

Effective test development requires a systematic, detail-oriented approach based on sound theoretical educational measurement principles. This chapter discusses twelve discrete test development procedures or steps that typically must be accomplished in the development of most achievement, ability, or skills tests. Following these twelve steps of effective test development, for both selected-response or constructed-response tests, tends to maximize validity evidence for the intended test score interpretation.

These twelve-steps are presented as a framework, which test developers may find useful in organizing their approach to the many tasks commonly associated with test development—starting with detailed planning in Step 1, carrying through to discussions of content definition and delineation, to creating test stimuli (items or prompts), and administering, scoring, reporting, and documenting all important test development activities. Relevant *Standards* are referenced for each step and validity issues are highlighted throughout.

This chapter provides an overview of the content of the *Handbook of Test Development*, with each of the twelve steps referenced to other chapters.

Effective test development requires a systematic, well-organized approach to ensure sufficient validity evidence to support the proposed inferences from the test scores. A myriad of details and issues, both large and small, comprise the enterprise usually associated with the terms *test development* and *test construction*. All of these details must be well executed to produce a test that estimates examinee achievement or ability fairly and consistently in the content domain purported to be measured by the test and to provide documented evidence in support of the

development activity receives sufficient attention to maximize the probability of creating an effective measure of the construct of interest.

This particular organization of tasks and activities into twelve discrete steps is somewhat arbitrary; these tasks could be organized differently such that there were fewer or more discrete steps. Each of these steps must be accomplished, at some level of detail, for all types of tests, whether the format is selected response (e.g., multiple choice), constructed response (e.g., short answer essay), or performance (e.g., high-fidelity simulation), and whatever the mode of test administration—traditional paper-and-pencil or computer based. The intensity and technical sophistication of each activity depends on the type of test under development, the test's purpose and its intended inferences, the stakes associated with the test scores, the resources and technical training of the test developers, and so on; but all of the tasks noted in this chapter must be carried out at some level of detail for every test development project.

These twelve steps provide a convenient organizational framework for collecting and reporting all sources of validity evidence for a testing program and also provide a convenient method of organizing a review of the relevant *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999) pertaining to test development. Each of these steps can be thought of as one major organizer of validity evidence to be documented in a technical report which summarizes all the important activities and results of the test. Each of these steps is also associated with one or more *Standards* (AERA, APA, NCME, 1999) that apply, more or less, given the exact purpose of the test, the consequences of test scores, and the desired interpretation or inferences from the test scores.

Table 1.1 lists the twelve steps of test development and provides a brief summary of tasks, activities, and issues; selected relevant *Standards* (AERA, APA, NCME, 1999) are also noted in Table 1.1 for each step. These steps are listed as a linear model or as a sequential timeline, from a discrete beginning to a final end point; however, in practice, many of these activities may occur simultaneously or the order of some of these steps may be modified. For example, if cut scores or passing scores are required for the test, systematic standard setting activities may occur (or a process may begin) much earlier in the test development process than Step 9 as shown in Table 1.1. Item banking issues are shown as Step 11, but for ongoing testing programs, many item banking issues occur much earlier in the test development sequence. But many of these activities are prerequisite to other activities; for example, content definition must occur before item development and test assembly, so the sequence of steps, although somewhat arbitrary, is meaningful.

Much of the information contained in this chapter is based on years of experience, learning what works and what does not work, in the actual day-to-day practice of test development. The relevant research literature supporting the best practice of test development comes from many diverse areas of psychometrics and educational measurement. The Millman and Greene chapter (1989) and the Schmeiser and Welch chapter (in press) in *Educational Measurement* inspire these 12-steps for effective and efficient test development.

STEP 1: OVERALL PLAN

Every testing program needs some type of overall plan. The first major decision is: What construct is to be measured? What score interpretations are desired? What test format or combination of formats (selected response or constructed response/performance) is most appropriate for the planned assessment? What test administration modality will be used (paper

TABLE 1.1
Twelve Steps for Effective Test Development

Steps	*Example Test Development Tasks*	*Example Related Standards*
1. Overall plan	Systematic guidance for all test development activities: construct; desired test interpretations; test format(s); major sources of validity evidence; clear purpose; desired inferences; psychometric model; timelines; security; quality control	Standard 1.1 Standard 3.2 Standard 3.9
2. Content definition	Sampling plan for domain/universe; various methods related to purpose of assessment; essential source of content-related validity evidence; delineation of construct	Standard 1.6 Standard 3.2 Standard 3.11 Standard 14.8
3. Test specifications	Operational definitions of content; framework for validity evidence related to systematic, defensible sampling of content domain; norm or criterion referenced; desired item characteristics	Standard 1.6 Standard 3.2 Standard 3.3 Standard 3.4 Standard 3.11
4. Item development	Development of effective stimuli; formats; validity evidence related to adherence to evidence-based principles; training of item writers, reviewers; effective item editing; CIV owing to flaws	Standard 3.6 Standard 3.7 Standard 3.17 Standard 7.2 Standard 13.18
5. Test design and assembly	Designing and creating test forms; selecting items for specified test forms; operational sampling by planned blueprint; pretesting considerations	Standard 3.7 Standard 3.8
6. Test production	Publishing activities; printing or CBT packaging; security issues; validity issues concerned with quality control	N/A
7. Test administration	Validity issues concerned with standardization; ADA issues; proctoring; security issues; timing issues	Standard 3.18 Standard 3.19 Standard 3.20 Standard 3.21
8. Scoring test responses	Validity issues: quality control; key validation; item analysis	Standard 3.6 Standard 3.22
9. Passing scores	Establishing defensible passing scores; relative vs. absolute; validity issues concerning cut scores; comparability of standards: maintaining constancy of score scale (equating, linking)	Standard 4.10 Standard 4.11 Standard 4.19 Standard 4.20 Standard 4.21
10. Reporting test results	Validity issues: accuracy, quality control; timely; meaningful; misuse issues; challenges; retakes	Standard 8.13 Standard 11.6 Standard 11.12 Standard 11.15 Standard 13.19 Standard 15.10

and pencil or computer based)? One needs to know how exactly and when to begin the program or process, in what sequence tasks must be accomplished, which tasks depend on the successful completion of other tasks, what timeline must be adhered to, who is responsible for carrying out which specific tasks, how to quality control all major aspects of the testing program, plus literally hundreds of other issues, decisions, tasks, and operational details. Step 1, the overall plan, places a systematic framework on all major activities associated with the test development project, makes explicit many of the most important a priori decisions, puts the entire project on a realistic timeline, and emphasizes test security and quality control issues from the outset.

Many of the most fundamental decisions about the testing program must be made prior to beginning formal test development activities. Each of these fundamental decisions, with its clear rationale, ultimately indicate a major source of validity evidence for the test scores resulting from the testing program.

Examples of Step 1–type tasks and decisions include a clear, concise, well-delineated purpose of the planned test. The purpose of testing forms an operational definition of the proposed test and guides nearly all other validity-related decisions related to test development activities. Ultimately, major steps such as content definition, the methods used to define the test content domain, and the construct hypothesized to be measured by the examination are all directly associated with the stated purpose of the test. The choice of psychometric model, whether classical measurement theory or item response theory, may relate to the proposed purpose of the test, as well as the proposed use of the test data and the technical sophistication of the test developers and test users. For example, if the clearly stated purpose of the test is to assess student achievement over a well-defined sequence of instruction or curriculum, the proposed construct, the methods used to select content to test, and the psychometric model to use are each reasonably clear choices for the test developer. Likewise, if the test's purpose is to estimate ability to select students for a national and highly competitive professional educational program, the inferences to be made from test scores are clearly stated, the major constructs of interest are delineated, and the content-defining methods, psychometric model, and other major test development decisions are well guided.

Many other fundamental decisions must be made as part of an overall test development plan, including: Who creates test items for selected-response tests or prompts or stimuli for performance tests? Who reviews newly written test items, prompts or other test stimuli? How is the item or prompt production process managed and on what timeline? Who is responsible for the final selection of test items or prompts? Who produces, publishes, or prints the test? How is test security maintained throughout the test development sequence? What quality controls are used to ensure accuracy of all testing materials? When and how is the test administered, and by whom? Is the test a traditional paper-and-pencil test or a computer-based test? If required, how is the cut score or passing score established, and by what method? Who scores the test and how are the scores reported to examinees? Who maintains an item bank or item pool of secure test items or performance prompts? What are the key dates on the timeline of test development to ensure that all major deadlines are met? Who is responsible for the complete documentation of all the important activities, data results, and evaluation of the test?

In many important ways, Step 1 is the most important step of the twelve tasks of test development. A project well begun is often a project well ended. This critical beginning stage of test development outlines all essential tasks to be accomplished for a successful testing project and clearly highlights all the important sources of validity evidence required for the testing program. Timelines and responsibilities are clearly stated, creating a reasonable and efficient plan to accomplish all necessary tasks, including allowing sufficient time for adequate quality control procedures and a correction cycle for all detected errors. Detailed, clear test program planning is an important first step toward adequately accomplishing the end goal of

preparing, administering, scoring, and analyzing a test and presenting reasonable sources of validity evidence to support or refute the intended inferences from test scores. All types of tests, used for whatever purpose, whether using traditional paper-and-pencil selected-response formats or performance tests using constructed-response formats or high-fidelity simulations benefit from the detailed planning activities of Step 1.

The *Standards* (AERA, APA, NCME, 1999) relating to the tasks of Step 1 discuss the importance of clearly defining the purpose of the test, following careful test development procedures, and providing a definitive rationale for the choice of psychometric model for scoring. For example, in discussing validity evidence, the *Standards* (AERA, APA, NCME, 1999, p. 17) suggest that ".... the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction...."

STEP 2: CONTENT DEFINITION

One of the most important questions to be answered in the earliest stages of test development is: What content is to be tested? All ability and achievement tests rely heavily on content-related validity evidence to make fundamental arguments to support (or refute) specific interpretations of test scores (Kane, chap. 7, this volume). No other issue is as critical, in the earliest stages of developing effective tests, as delineating the content domain to be sampled by the examination. If the content domain is ill defined or not carefully delineated, no amount of care taken with other test development activities can compensate for this inadequacy. The validity of inferences for achievement test scores rests primarily and solidly on the adequacy and defensibility of the methods used to define the content domain operationally, delineate clearly the construct to be measured, and successfully implement procedures to systematically and adequately sample the content domain.

The chapter on practice analysis (Raymond & Neustel, chap. 9, this volume) presents a thorough discussion of this empirical method of content definition, especially in the context of credentialing examinations. The *Standards* (AERA, APA, NCME, 1999) clearly endorse the use of empirical job analysis for selection and employment examinations (Standard 14.8), but leave the methodology for content definition more open-ended for other types of examinations. Webb (chap. 8, this volume) addresses the defining of content in achievement tests.

Content defining methods vary in rigor, depending on the purpose of the test, the consequences of decisions made from the resulting test scores, and the amount of defensibility required for any decisions resulting from test scores. For some lower stakes achievement tests, the content-defining methods may be very simple and straightforward, such as instructors making informal (but informed) judgments about the appropriate content to test. For other very high-stakes examination programs, content definition may begin with a multiyear task or job analysis, costing millions of dollars, and requiring the professional services of many testing professionals.

For high-stakes achievement examinations, test content defining methods must be systematic, comprehensive, and defensible. For instance, a professional school may wish to develop an end-of-curriculum comprehensive achievement test covering the content of a two-year curriculum, with a passing score on this test required to continue on to a third year of professional education. In this example, rigorous and defensible methods of content definition and delineation of the content domain are required; all decisions on content, formats, and methods of content selection become essential aspects of validity evidence. In this situation, faculty and administrators may need to develop systematic, thorough methods to define all the content "at risk" for such a high-stakes test, looking to curricular documents, teaching

syllabi, instructional materials and content, textbook content, and faculty judgment to behaviorally and operationally define content. The input of professional measurement specialists is desirable for such a high-stakes testing environment. Instructors might be asked to complete empirical methods similar to practice or job analysis to rate and rank "the importance to test" of content statements, after comprehensive lists of all content had first been identified. Their judgments might be further refined by asking representative faculty to rate the "criticality" of content statements or make some judgments about how essential the content statement is to future learning or to the ultimate successful practice of the professional discipline. From such a multistage content-defining process, a complete content domain could result, from which a sampling plan to guide the creation and selection of item content could then be developed.

Credentialing tests in the professions (or occupations) are often required as one of several sources of evidence of competence to be licensed to practice the profession in a jurisdiction. Public safety is typically the greatest concern for licensing tests, such that the persons using the services of the professional have some minimal confidence in the professional's ability to deliver services safely. Clearly, the content-defining and delineation methods used for such high-stakes tests must be much more rigorous and defensible than for many (if not most) other testing situations. Formal practice or task analysis methods (Raymond & Neustel, chap. 9, this volume) are usually required for these high-stakes tests, because the consequences of misclassifications can have serious impact on both society and the individual professional seeking a license to practice.

Certification examinations in the professions share many of the same content-defining requirements as licensening examinations noted above. Traditionally, certification was thought to be voluntary and "over and above" a basic license to practice. More recently, the distinction between licensure and certification has been blurred in many professions. In some professions, like medicine, certification in a medical specialty (such as General Surgery) is essentially required if the individual hopes to practice in the specialty or subspecialty. The content-defining methods required to support (or refute) validity arguments and support the ultimate defensibility for many such certifying examination programs must be extremely rigorous (Raymond & Neustel, chap. 9, this volume).

The defensibility of the content-defining process is associated with the rigor of the methods employed. One essential feature is the unbiased nature of the methods used to place limits around the universe or domain of knowledge or performance. Further, the requirements for defensibility and rigor of content-defining methods is directly proportional to the stakes of the examination and the consequences of decisions made about individuals from the resulting test scores. Rigor, in this context, may be associated with attributes such as dependability of judgments by content experts or subject matter experts (SMEs), qualifications of the SMEs making the content judgments, the inherent adequacy of the judgmental methods employed (lack of bias), the number, independence, and representativeness of the SMEs, and so on. Empirical methods may be used to support the adequacy of the judgmental methods. For example, it may be important to demonstrate the reproducibility of independent SME judgments by computing and evaluating a generalizability coefficient (Brennan, 2001) for raters or some other statistical index of consistency of ratings.

Methods chosen to define test content are critical and depend on the purpose of the test, the consequences of the decisions made on the basis of the test scores, and the validity evidence needed to support test score interpretations. The definition of test content is ultimately a matter of human judgment. Methods and procedures may be developed and used to minimize bias and increase the objectivity of the judgments, but—in the end—professional judgments by content experts shape the content domain and its definition. Such inherent subjectivity can lead to controversy and disagreement, based on politics rather than content.

STEP 3: TEST SPECIFICATIONS: BLUEPRINTING THE TEST

The process of creating test specifications guides detailed test development activities and completes the operational planning for tests in a systematic manner. *Test specifications* and *test blueprint* are sometimes used almost interchangeably. For this chapter, *test specifications* refers to a complete operational definition of test characteristics, in every major detail, and thus includes what some authors call the *test blueprint*. For example, at a minimum, the test specifications must describe (1) the type of testing format to be used (selected response or constructed response/performance); (2) the total number of test items (or performance prompts) to be created or selected for the test, as well as the type or format of test items (e.g., multiple choice, three option, single-best answer); (3) the cognitive classification system to be used (e.g., modified Bloom's taxonomy with three levels); (4) whether or not the test items or performance prompts will contain visual stimuli (e.g., photographs, graphs, charts); (5) the expected item scoring rules (e.g., 1 point for correct, 0 points for incorrect, with no formula scoring); (6) how test scores will be interpreted (e.g., norm or criterion referenced); and (7) the time limit for each item. A *test blueprint* defines and precisely outlines the number (or proportion) of test questions to be allocated to each major and minor content area and how many (what proportion) of these questions will be designed to assess specific cognitive knowledge levels. The higher the stakes or consequences of the test scores, the greater detail should be associated with the test specifications; detailed test specifications provide a major source of validity evidence for the test. But, at minimum, all test specifications must include some range of expected items/prompts to be selected for each major content category and each major cognitive process level (Linn, chap. 2, this volume).

The test specifications form an exact sampling plan for the content domain defined in Step 2. These documents and their rationales form a solid foundation for all systematic test development activities and for the content-related validity evidence needed to support score inferences to the domain of knowledge or performance and the meaningful interpretation of test scores with respect to the construct of interest.

How do the content defining activities of Step 2 get translated into exact test specifications in Step 3? The rigor and sophistication of the blueprinting methods will depend on the consequences of testing. Table 1.2 presents a simple example of a test blueprint for an achievement examination over a curriculum on test development. This simple blueprint operationalizes judgments about the content appropriate for sampling in this achievement test. In this example, the test developer has allocated the item content in proportion to the amount of instructional time devoted to each topic and has further weighted these allocations using some professional judgment about the relative importance of topics. For example, content dealing with "item

TABLE 1.2
Example Test Blueprint—Achievement Examination on Test Development

Content Area	*Recall*	*Application*	*Problem Solving*	*Totals*
Content define	4	10	6	20
Test specs	3	8	4	15
Item writing	4	10	6	20
Assembly: print administer	2	5	3	10
Test scoring	3	7	5	15
Test standards	4	10	6	20
Totals	20%	50%	30%	100%

writing" was judged to be about twice as important as content dealing with "test assembly, printing, and administration." Such judgments are subjective, but should reflect the relative emphasis in the curriculum as represented by the instructional objectives. This blueprint further suggests that the test developer has some judgmental basis for allocating one half of all test questions to the cognitive level labeled "application," whereas only 20 percent is targeted for "recall" and 30 percent for higher order "problem solving." These cognitive level judgments must reflect the instructional objectives and the instructional methods used for teaching and learning.

For high-stakes, large-scale examinations such as licensing or certifying examinations, the methods used to operationalize the task or practice analysis are much more formal (e.g., Raymond, 2001; Raymond & Neustel, chap. 9, this volume; Spray & Huang, 2000). Webb's chapter (Webb, chap. 8, this volume) addresses many of these issues for other types of achievement tests, especially those used for high-stakes accountability tests in the schools. Typically, the detailed results of an empirical practice analysis are translated to test specifications using some combined empirical and rational/judgmental method. For example, the empirical results of the practice analysis may be summarized for groups of representative content experts, who are tasked with using their expert judgment to decide which content areas receive what specific relative weighting in the test, given their expert experience.

The *Standards* (AERA, APA, NCME, 1999) emphasize documentation of the methods used to establish test specifications and blueprints, their rationale, and the evidence the test developers present to support the argument that the particular test specification fairly represents the content domain of interest. The *Standards* relating to test specifications and test blueprints emphasize the use of unbiased, systematic, well-documented methods to create test sampling plans, such that the resulting examination can have a reasonable chance of fairly representing the universe of content.

STEP 4: ITEM DEVELOPMENT

This step concentrates on a discussion of methods used to systematically develop selected-response items, using the multiple-choice item form as the primary exemplar. Downing (chap. 12, this volume) discusses the selected-response formats in detail in another chapter and Welch (chap. 13, this volume) discusses development of performance prompts.

Creating effective test items may be more art than science, although there is a solid scientific basis for many of the well-established principles of item writing (Haladyna, Downing, & Rodriguez, 2002). The creation and production of effective test questions, designed to measure important content at an appropriate cognitive level, is one of the greater challenges for test developers.

Early in the test development process, the test developer must decide what test item formats to use for the proposed examination. For most large-scale, cognitive achievement testing programs, the choice of an objectively scorable item format is almost automatic. The multiple-choice format (and its variants), with some ninety years of effective use and an extensive research basis, is the item format of choice for most testing programs (Haladyna, 2004). Test developers need not apologize for using multiple-choice formats on achievement tests; there is strong research evidence demonstrating the high positive correlation between constructed-response and selected-response item scores for measuring knowledge and many other cognitive skills (Rodriguez, 2003). (To measure complex constructs, such as writing ability, a constructed response format is typically required.)

The multiple-choice item is the workhorse of the testing enterprise, for very good reasons. The multiple-choice item is an extremely versatile test item form; it can be used to test all levels of the cognitive taxonomy, including very high-level cognitive processes (Downing, 2002a).

The multiple-choice item is an extremely efficient format for examinees, but is often a challenge for the item writer.

The choice of item format is a major source of validity evidence for the test. A clear rationale for item format selection is required. In practice, the choice of item form—selected response versus constructed response—may quite legitimately rest largely on pragmatic reasons and issues of feasibility. For example, for a large-scale, paper-and-pencil examination program, it may not be cost effective or time efficient to use large numbers of constructed response questions. And, given the research basis supporting the use of multiple choice items (e.g., Downing, 2002a, 2004; Haladyna, 2004; Rodriguez, 2003), the test developer need not feel insecure about the choice of a low-fidelity selected-response format, like the multiple choice format, for an achievement test.

The principles of writing effective, objectively scored multiple-choice items are well established and many of these principles have a solid basis in the research literature (Downing, 2002b, 2004; Haladyna, 2004; Haladyna & Downing, 1989a,b; Haladyna et al., 2002). Yet, knowing the principles of effective item writing is no guarantee of an item writer's ability to actually produce effective test questions. Knowing is not necessarily doing. Thus, one of the more important validity issues associated with test development concerns the selection and training of item writers (Abedi, chap. 17, this volume; Downing & Haladyna, 1997). For large-scale examinations, many item writers are often used to produce the large number of questions required for the testing program. The most essential characteristic of an effective item writer is content expertise. Writing ability is also a trait closely associated with the best and most creative item writers. For some national testing programs, many other item writer characteristics such as regional geographic balance, content subspecialization, and racial, ethnic, and gender balance must also be considered in the selection of item writers. All of these item writer characteristics and traits are sources of validity evidence for the testing program and must be well documented in the technical report (Becker & Pomplun, chap. 30, this volume).

Item Writer Training

Effective item writers are trained, not born. Training of item writers is an important validity issue for test development. Without specific training, most novice item writers tend to create poor-quality, flawed, low-cognitive-level test questions that test unimportant or trivial content. Although item writers must be expert in their own disciplines, there is no reason to believe that their subject matter expertise generalizes to effective item writing expertise. Effective item writing is a unique skill and must be learned and practiced. For new item writers, it is often helpful and important to provide specific instruction using an item writer's guide, paired with a hands-on training workshop (Haladyna, 2004). As with all skill learning, feedback from expert item writers and peers is required. The instruction–practice–feedback–reinforcement loop is important for the effective development and maintenance of solid item writing skills (Jozefowicz et al., 2002).

Competent content review and professional editing of test questions is also an important

in high-stakes tests (e.g., Baranowski, chap. 15, this volume; Abedi, chap. 17, this volume). Professional test editors often find content errors or inconsistencies, which can be clarified by item authors or item content reviewers. Additional content reviews are required, by independent subject matter experts, for high-stakes examinations. Such independent content reviews, after professional editorial review, strengthen the content-related validity evidence for the test.

All test item writers benefit from specialized training. Some of the worst test item examples are found on instructor-developed tests, at all levels of education (e.g., Mehrens & Lehmann, 1991). Poor quality and flawed test items introduce construct-irrelevant variance (CIV) to the assessment, potentially decreasing student passing rates by increasing the mean item difficulty (Downing, 2002b, 2004). Ideally, all item writers with any responsibility for achievement assessment have some special expertise in item writing and test development, gained through effective training and practice. Unfortunately, this is too often not the case and is one of the major failings of educational assessment at all levels of education, from K–12 to graduate professional education.

The creation of effective test items (or performance prompts) is a challenging but essential step in test development. Because the test item is the major building block for all tests, the methods and procedures used to produce effective test items is a major source of validity evidence for all testing programs. This fact is reflected in at least five separate standards with respect to test items and their creation and production.

STEP 5: TEST DESIGN AND ASSEMBLY

Assembling a collection of test items (or performance prompts) into a test or test form is a critical step in test development. Often considered a mundane task, the validity of the final test score interpretation very much relies on the competent and accurate test assembly process. Quality control is the keyword most associated with test assembly. The absence of errors in the test assembly process often goes unnoticed; errors or serious flaws and omissions in the test assembly process can be obviously glaring and have the potential of seriously reducing the validity evidence for examination scores.

The overall design of the test, planned in detail in Step 1, provides a sound rationale related to the purpose of testing and the planned interpretation and use of the test scores. This formal overall test design creates the theoretical basis for Step 5. Several other chapters in this book address test design issues in detail (see Davey & Pitoniak, chap. 24; Luecht, chap. 25; Roid, chap. 23; Jones, Smith, & Talley, chap. 22; Wendler & Walker, chap. 20; Young, chap. 21, this volume).

The specific method and process of assembling test items into final test forms depends on the mode of examination delivery. If a single test form is to be administered in paper-and-pencil mode, the test can be assembled manually, by skilled test developers (perhaps using computer software to assist manual item selection and assembly). If multiple "parallel" test forms are to be assembled simultaneously, human test developers using advanced computer software can assemble the tests. If the test is to be administered as computer-based, more specialized computer software will likely be needed to assemble multiple test forms to ensure proper formatting of the fixed-length test form for the computer-delivery software. If the test is to be administered as a computer-adaptive test, very advanced computer software (automatic test assembly software) will likely be required to adequately manage and resolve multiple test item characteristics simultaneously to create many equivalent test forms (Luecht, chap. 25, this volume).

For most achievement tests, whatever the delivery mode, the most important validity-related issues in test assembly are the correspondence of the content actually tested to the content

specifications developed in Step 3 and the high-level quality control of this entire process. The test assembly step operationalizes the exacting sampling plan developed in Steps 2 and 3 and lays the solid foundation for inferential arguments relating sample test scores to population or universe scores in the domain. This is the essence of the content-related validity argument, which, to be taken seriously, must be independently verifiable by independent, noninvested content experts (see Kane, chap. 7, this volume, for a complete discussion).

Other major considerations in test assembly, at least for traditional paper-and-pencil tests, relate to formatting issues (see Campion & Miller, chap. 26, this volume, for a complete discussion). Tests must be formatted to maximize the ease of reading and minimize any additional cognitive burden that is unrelated to the construct being tested (e.g., minimize CIV). Traditional principles such as formatting items such that the entire item, together with any visual or graphical stimuli, appear on the same page (or frame) falls under this rubric of minimizing the potential CIV associated with overly complex item formatting.

Other formatting issues are more psychometric in nature. For example, the placement of pretest (tryout) items within the test form is a formatting issue of interest and concern. Ideally, pretest items are scattered throughout the test form randomly, to minimize any effects of fatigue or lack of motivation (if items are recognized by examinees as pretest only items). (Such random distribution of pretest items is not always practical or feasible.) Both *Standards* cited as relating to Step 5 address pretest or tryout items, suggesting thorough documentation of the item sampling plan and the examinee sampling plan.

Key Balance of Options and Other Issues

Balance of the position or location of the correct answer is an important principle for all selected-response items assembled into discrete test forms. The principle is straightforward: There should be an approximately equal frequency of correct responses allocated to the first position (e.g., *A*), and the second position (*B*), and so on. Approximating this goal is sometimes difficult, given other important constraints from the item writing principles. For example, all options containing numbers must be ranked from high to low or low to high, such that the location of the correct answer cannot be rearranged in such an item. The reason for this key balance principle is that both item writers and examinees have a bias toward the middle position, such that item writers tend to place the correct answer in the middle position and examinees tend to select a middle-position answer as a natural testwise inclination (Attali & Bar-Hillel, 2003).

The placement of *anchor* or *common* items used in a common-item equating design is also a formatting issue with validity implications (Kolen & Brennan, 2004). Ideally, the common items used to link test forms in a classical equating design should appear in a similar location in the new test as they did in the prior test to eliminate any potential "order effect" on the common items' difficulty and/or discrimination. (Because of other operational and practical considerations, it is not always possible to place anchor equating items in exactly the same location.)

of test assembly will occur and these errors will reduce the legitimacy of the final test score interpretation, thus reducing the validity evidence for the scores.

Although many of the details of assembling tests into final forms may seem routine and mundane, the product of Step 5 is the test that examinees will encounter. Test assembly errors reduce validity evidence by introducing systematic error—CIV—to the test (Haladyna & Downing, 2004) and may lead to the invalidation of scores for some test items, which potentially reduces the content-related validity evidence for the test.

STEP 6: TEST PRODUCTION

The production, printing, or publication of examinations is another routine step of test development that is often overlooked with respect to its validity aspects. For example, there appear to be no *Standards* that bear directly on this important test development activity. All the prior test development work comes to fruition in Step 6, where the months or years of prior test development work is finally "cast in stone." Campion and Miller (chap. 26, this volume) discuss test production issues and their effect on test validity issues in detail.

All tests, whether large-scale national testing programs or much smaller local testing programs, must ultimately be printed, packaged for computer administration, or published in some form or medium. Test production activities and their validity implications apply equally to performance tests and selected-response tests. Step 5—Test Production—truly operationalizes the examination, making final all test items, their order, and any visual stimuli associated with the test items. This is "the test," as it will be experienced by the examinee, for better or worse. Clearly, what the examinee experiences as the final test form has major implications for how scores on the test can be interpreted, and this is the critical validity aspect associated with test production.

Security issues are prominent for test production. Human error is the most likely source of test security breaches, even in this era of high-tech computer-assisted test production.

During the production process, whether it be physical test booklet printing or packaging activities for computer-based tests, final test items may be available in some form to more individuals that at any prior time during test development. All reasonable security precautions must be taken during test production, during the electronic transmission of secure test items, secure shipping of printed test copy and printed booklets, and secure destruction of excess secure printed materials. Test production security standards and policies must be developed, implemented, and quality controlled for all high-stakes examinations and should reflect the consequences of testing somewhat proportionally. Independent audits of these security procedures should be carried out periodically, by security professionals, especially for all computer-based security systems. All secure test materials must be locked in limited-access files at all times not in direct use by test developers and production staff. For high-stakes tests, staff should have access only to those materials needed for completing their specified tasks and on a time-limited basis; high security control must be maintained and frequently reviewed and updated for all computer systems (Impara & Foster, chap. 5, this volume).

For printed tests, printers usually can provide some type of off-press copy for final review by test development staff. This final preprinting quality control step is important even for tests that are printed directly from camera-ready copy or from a direct electronic file; typographical errors or other major potential item invalidating errors, which were missed at all other proofreading and quality control stages, can often be identified (and corrected) at this late stage.

Other quality control issues are equally important for test production. For example, if a test is being printed by a printing company or service, test development staff (in addition to

printer's staff) must take responsibility for many quality assurance procedures. This may mean that test development staff randomly sample some number of final printed booklets to ensure the completeness of test booklets (e.g., no pages missing) and the overall quality (readability) of the final printed copy, including visual material, and so on.

The quality and readability of final test printing or production is important to the validity evidence for the test (Campion & Miller, chap. 26, this volume). If test items or visual stimuli are unclearly printed, test booklet pages are omitted, items are misordered, or options are out of order, such items are potentially invalidated and must be omitted from final scoring. Production and printing errors (or formatting errors for computer-based tests) can seriously reduce the validity evidence for a test and create an aura of distrust and anxiety concerning many other aspects of the test, its construction, and scoring. Great care and effective quality control measures must be exercised during the final production process for all tests, no matter what the modality of delivery. Maintaining complete control over test security, with independently verifiable audit trails, is essential during the production process.

STEP 7: TEST ADMINISTRATION

The administration of tests is the most public and visible aspect of testing. There are major validity issues associated with test administration because much of the standardization of testing conditions relates to the quality of test administration. Whether the test is administered in local school settings by teachers, in large multisite venues by professional proctors, or by trained staff at nationwide computer-based testing centers, many of the basic practices of sound test administration are the same (see McCallin, chap. 27, this volume, for a detailed discussion of test administration issues and Thurlow, Thompson, & Lazarus, chap. 28, this volume, for a discussion of special needs test administration issues, including Americans With Disabilities Act issues.)

Standardization is a common method of experimental control for all tests. Every test (and each question or stimulus within each test) can be considered a mini experiment (van der Linden & Hambleton, 1997). The test administration conditions—standard time limits, proctoring to ensure no irregularities, environmental conditions conducive to test taking, and so on—all seek to control extraneous variables in the "experiment" and make conditions uniform and identical for all examinees. Without adequate control of all relevant variables affecting test performance, it would be difficult to interpret examinee test scores uniformly and meaningfully. This is the essence of the validity issue for test administration considerations.

Security is a major concern for test administration. For most examinations (e.g., large-scale, high-stakes tests), the entire test development process is highly secure, with extremely limited access to test materials. Much effort and expense are devoted to securing examination items and test forms. This "chain of security" is necessarily widened during test production and becomes extremely wide for large-scale test administration. For paper-and-pencil examinations, which are administered in multiple sites, printed test forms and all testing materials must be securely shipped to test sites; securely received and maintained by proctors; distributed to examinees

including supervision of other proctors on site (see McCallin, chap. 27, this volume, for detailed discussion of proctoring).

For large-scale computer-based tests, proctoring is generally delegated to the agency providing the computer test administration network. Computer-based testing changes the test security environment. Some security issues associated with paper-and-pencil testing are eliminated, such as printing of test booklets, shipping secure test materials, distributing printed test forms to examinees, and so on. However, other potential test security vulnerabilities are increased, such as the electronic transmission of secure test items and response data, the need to have large numbers of items available on local computer servers, and the possible use of less well-trained and less professional administrative staff in the testing sites.

The *Standards* associated with test administration generally deal with issues of standardization and examinee fairness, such as time limits, clarity of directions to examinees, and standard conditions for testing.

Test administration is an extremely important component of test development. Competent, efficient, and standardized administration of tests provides important validity evidence for test scores. Deficiencies in the detailed planning and logistics required for high-quality test administration can lead to a serious reduction in validity evidence for the examination. The importance of test administration to validity evidence is equally true for a selected-response test as it is for a performance examination; a performance test typically adds many challenges and complex logistics for administration. Security problems during test administration can lead to the invalidation of some or all examinee scores and can require the test developers to retire or eliminate large numbers of very expensive test items. Improper handling of issues pertaining to the Americans with Disabilities Act (ADA) can lead to legal challenges for the test developers. Improper overuse of ADA accommodations can lead to misinterpretations of test scores and other types of validity issues associated with score misinterpretation.

STEP 8: SCORING EXAMINATION RESPONSES

Test scoring is the process of applying a scoring key to examinee responses to the test stimuli. Examinee responses to test item stimuli or performance prompts provide the opportunity to create measurement. The responses are not the measurement; rather an application of some scoring rules, algorithms, or rubrics to the responses result in measurement (Thissen & Wainer, 2001; van der Linden & Hambleton, 1997).

There are many fundamental validity issues associated with test scoring. The most obvious issue relates to accuracy of scoring. If final test scores are to have valid meaning, especially the meaning that was anticipated by the test developers (a measure of the construct of interest, an adequate sample of the domain of knowledge, and so on), a scoring key must be applied with perfect accuracy to the examinee item responses. Scoring errors always reduce validity evidence for the test and can invalidate the results. Validity evidence can be reduced by either a faulty (inaccurate) scoring key or flawed or inaccurate application of the scoring key to responses. Thus, high levels of quality control of the scoring process are essential to validity.

Scoring can be extremely simple or very complex, depending on the type of test item or stimuli. The responses to single-best-answer multiple choice items are easily scored by computer software, whereas responses to complex computer simulation problems can be more challenging to score reliably. Selected-response items are generally more efficiently and objectively scored than constructed-response items and performance prompts; however, constructed-response and performance items can be scored accurately and reliably by competently trained and monitored scorers or computer software.

All scoring issues, for all types of tests and testing formats—from relatively straightforward selected-response formats to extremely complex performance simulations—concern the accurate representation of examinee performance with respect to the measured construct or the domain of knowledge, skills, or ability (e.g., Thissen & Wainer, 2001). Ideally, the scored examination responses correspond closely (nearly one to one) with the examinee's true state with respect to the domain measured by the test or construct of interest. Insofar as the test scores depart from this one-to-one correspondence, random measurement error or systematic construct-irrelevant measurement error (CIV) has been introduced into the measurement.

Much of psychometric theory begins at the "scored response" point of test development. In Step 8, a basic test scoring process, which is appropriate for nearly all achievement tests, is discussed.

Preliminary Scoring and Key Validation

A preliminary scoring and item analysis with a final verification of the scoring key by content experts is an essential quality control procedure for many tests (see Livingston, chap. 19, this volume, for a complete discussion of item analysis issues).

A final "key validation" or key verification step increases the validity evidence for all examinations. This two-step scoring process is essential for tests containing newly written and non-pretested items, because it is possible that such items may contain invalidating flaws that were not detected during the item writing and review process. *Key validation* is the process of preliminary scoring and item analysis of the test data, followed by a careful evaluation of the item-level data to identify potentially flawed or incorrect items prior to final test scoring (Downing & Haladyna, 1997). This key validation process is nearly identical for selected-response tests and performance tests and should be carried out for all types of test modalities, if possible (e.g., Boulet, McKinley, Whelan, & Hambleton, 2003).

After examinee item responses are scanned from paper-and-pencil answer sheets, response strings are provided from computer-based testing administration, or data are computer entered for performance examinations, an initial scoring and item analysis should be completed. Items that perform anomalously should be identified for final review of the scoring key or other potential content problems by subject matter experts. Item difficulty and item discrimination criteria or ranges for key validation item identification should be developed for every testing program (see Haladyna, 2004, p. 228, for example). Typically, items that are very difficult and/or have very low or negative item discrimination indices are identified for further content review. The key validation criteria should be sufficiently sensitive that all potentially problematic questions are identified for final content review. The results of key validation, such as the number and type of questions identified for review, and the disposition of these items (scored "as is," eliminated from final scoring, or key changed) are a source of validity evidence and should be documented in the technical report.

If large numbers of items are identified for key validation procedures, the criteria are either too liberal or there are serious problems with the item writing and review process. If large numbers of items are eliminated from the final scoring, for reasons of poor item quality or incorrectness of content, content-related validity evidence is compromised.

Final Scoring

Final scoring of the examinee responses follows the preliminary scoring and key validation procedures. The final answer key must be carefully proofread and quality controlled for absolute accuracy. Great care must be taken to ensure the complete accuracy of this final scoring key. Subject matter experts who have direct responsibility for the content of the examination must

formally approve the final scoring key, especially if testing services are being provided by some outside or contract agency. If multiple forms of the same examination are used, great care must be taken to ensure the use of the correct scoring key for each test form. All pretest or tryout items must be carefully segregated from the final scoring so that pretest questions do not contribute to the final test score in any way (see Livingston, chap. 19, this volume).

A final item analysis should be completed and reviewed carefully. The final item analysis provides another important quality control step to ensure that any changes made at the key validation stage are accurately reflected in the final scoring. A complete final item analysis includes summary test statistics for the test administration. For tests using classical measurement theory, these statistics include the raw score mean and standard deviation, the mean item difficulty (p-value), mean item discrimination (point-biserial or biserial), range of raw scores, the test score reliability (Alpha/Kuder-Richardson 20), plus other appropriate indices of overall test quality such as some index of pass–fail reproducibility (especially for high-stakes examinations). Summary test statistics are critically important validity evidence and must be thoroughly evaluated and documented. Any anomalies identified by final item analysis or final summary test statistical analyses must be thoroughly investigated and resolved prior to reporting test scores.

If test score scaling or equating procedures are to be carried out, these procedures usually follow the final scoring and item analysis (unless pre-equating methods are used).

The standards related to scoring issues discuss ensuring the correspondence of the scoring rules to the stated purpose of testing and clarity of scoring rules to ensure absolute accuracy of final scores.

The most important emphasis in the test scoring step is complete accuracy. Extreme quality control procedures are required to ensure total accuracy of final test scores, especially for very high-stakes examinations. Any scoring errors included in final test scores reduces validity evidence and credibility of the examination and introduces CIV to scores (Haladyna & Downing, 2004)

STEP 9: ESTABLISHING PASSING SCORES

Many, but not all, tests require some type of cut score (passing score) or performance standard. For tests that require cut scores, content standard interpretations, or "grade levels" attributed to certain test scores or score ranges, the methods and procedures used to establish cut scores are a major source of validity evidence and are an integral part of the test development process (see Cizek, chap. 10, this volume, for a complete discussion of standard setting).

For high-stakes tests, the establishment of defensible cut scores is one of the most critical test development issues. For tests of all types, with any consequences for examinees, the legitimacy of the methods used to identify cut scores is a major source of validity evidence.

Step 9 highlights and generally overviews some of the important issues concerning standard setting. Standard setting is a complex issue with a sound basis in the research literature. Also, it must be noted that methods and procedures used to establish passing scores may take place at several different stages of test development, depending on the methods used and the overarching philosophy of standard setting adopted by the test developers and users. The basic decisions on the type of standard-setting method to use should be made early in the test development process, because extensive planning may be required to successfully implement certain types of standard setting procedures. Further, some methods may require multiple exercises or studies, taking place at different times during the test development process and concluding after the test has been administered.

Relative and Absolute Standard-Setting Methods

All methods of establishing passing scores require human judgment and are therefore somewhat arbitrary (e.g., Norcini & Shea, 1997). All examination passing scores answer the question: How much knowledge (skill or ability) is needed to be classified as having passed the examination? Traditionally, standard-setting methods are dichotomized into two major categories: relative or normative methods and absolute methods. But, there are standard-setting methods that blend characteristics of both relative and normative methods, such as the relative–absolute or Hofstee method (Hofstee, 1983).

Relative standard-setting methods—normative methods—use the actual test performance data, usually from some well-defined group of test takers (e.g., first-time takers of the test) to establish a point on the distribution of test scores to designate as the cut score or passing point (Cizek, 2001). The passing score point on the distribution of scores represents a judgment of someone or some group of qualified individuals who are responsible for making such judgments. For example, relative passing scores might be expressed as a score that is exactly one standard deviation below the mean score of all examinees; a T-score of 45, based on all examinees who took the examination for the first time; or a percentile rank of 30. The emphasis of all normative passing scores is on the relative position of the examinee's score in some distribution of scores. All examinees scoring at or above the selected cut point on the distribution pass the test and all those scoring below that point fail. (Note that it is the method used to establish the passing score that makes the chosen score "relative," not the score metric.) Relative passing scores do not address the absolute competency of examinees or make any judgments about what specific knowledge, skill, or ability has been mastered by the examinee.

Absolute passing score methods employ systematic procedures to elicit expert judgments from SMEs, concerning the amount of knowledge (skill or ability) required on a test to be considered a passing examinee. There are many well-established, well-researched methods commonly used to establish defensible and effective absolute passing scores (Cizek, 2001; Cizek, chap. 10, this volume).

All common methods used to establish absolute passing scores on all types of examinations require expert judgments about the expected performance of a borderline examinee. A *borderline examinee* is usually defined as someone who just barely passes or just barely fails the test; the borderline examinee has an exactly equal probability (50–50) of either passing or failing the test. Some commonly used methods are the Angoff method and its modifications (Angoff, 1971; Impara & Plake, 1997; Downing, Lieska, & Raible, 2003), which requires expected passing score judgments about each individual test question; the Ebel method (Ebel, 1972), which requires judges to make expected passing score judgments for sets of items that have been classified into difficulty and relevance categories; and the Hofstee method (Hofstee, 1983), with both absolute and relative characteristics, which asks judges to state their own expected minimum and maximum passing score and failure rate (proportion of examinees failing the test) on the test and then plots those expert judgments onto actual test data.

The absolute methods are not without their critics. For example, Zieky (1997) suggests that the judgments required of content experts, regarding the expected pass score for borderline examinees, represents an impossible cognitive task. Other controversies concern the amount of performance data to provide the content expert judges, although measurement experts tend toward providing more rather than less empirical performance data to standard setting judges (e.g., Zieky, 2001).

Other, more data-centered methods, such as the contrasting groups method and the borderline groups method, are also used for certain types of performance examinations. In the *contrasting groups method*, the actual test performance of a known group of masters, experts,

or highly qualified examinees is plotted against the actual test performance of a known group of nonmasters or those who are known to be non-expert or not qualified. The intersection of the two curves describes the passing score, which can be adjusted to minimize false-positive or false-negative error. The *borderline group method* is similar, but requires a direct expert judgment concerning which examinees are "borderline" in their performance; these judgments are then translated into passing scores on the examination (e.g., Wilkinson, Newble, & Frampton, 2001; Kilminster & Roberts, 2003).

The comparability of passing scores across different forms (different test administrations) is a major validity issue (e.g., Norcini & Shea, 1997). If absolute passing scores are used, it is critical that test score equating be used to maintain the constancy of the score scale. If scores are not equated, even slight differences in mean item difficulty across different test administrations make the interpretation of the passing score impossible and may unfairly advantage or disadvantage some examinees (Kolen & Brennan, 2004). Thus, most of the *Standards* (AERA, APA, NCME, 1999) concerning passing scores discuss equating issues. Other standards for absolute passing score determination address issues of ensuring that the task of the standard-setting judges is clear and that the judges can in fact make reasonable and adequate judgments. Fairness of cut score procedures and attention to the consequences or the impact of passing scores are emphasized by the *Standards* (AERA, APA, NCME, 1999).

In summary, different standard-setting methods, whether the traditional relative methods or the more contemporary absolute methods (or other blended methods), produce different cut scores and passing rates. None of these methods is more correct than other methods. The task of content-expert judges, using absolute standard-setting methods, is not to discover some true passing score, but rather to exercise their best professional judgment in answering the question: How much is enough (to pass)? Passing scores reflect policy, values, expert judgment, and politics. The defensibility and the strength of the validity evidence for passing scores relies on the reasonableness of the unbiased process, its rationale and research basis, and the psychometric characteristics of expert judgments.

STEP 10: REPORTING EXAMINATION RESULTS

Score reporting is an important, often complex, step in test development. The contents and format of an examinee score report should be among the many early decisions made for large-scale testing programs. There are multiple validity issues concerning score reporting and several *Standards* (AERA, APA, NCME, 1999) address the adequacy of requirements for score reporting. Ryan (chap. 29, this volume) discusses elements of score reports and the strategies behind different types of score reports.

For large-scale assessments, the score reporting task is complex (e.g., Goodman & Hambleton, 2004) and is often encumbered with nonpsychometric issues. The score reporting issues emphasized by the relevant *Standards* deal with fairness, timeliness, appropriateness of the score, avoidance of score misunderstanding and misuse, and tangentially, issues of test retake (for failing examinees), and test score challenges.

As with so many other prior steps of test development, absolute accuracy is of the highest importance for all reports of scores. Thus, careful and effective quality control measures are critically important. For large-scale, high-stakes examinations, one of the most catastrophic errors is to publicly distribute incorrect score reports, particularly if the pass–fail status of some examinees changes as a result of the scoring errors. Even somewhat trivial errors associated with score reports, such as typographical errors or formatting errors, can call into question the accuracy of the reported score and degrade the credibility of the entire testing program.

For some types of high-stakes examinations, such as licensure and certification examinations, only the pass–fail results of examinations may be reported to examinees. This practice is considered minimally acceptable. However, many high-stakes testing programs report total test scores, the passing score, and some relevant and informative subscale scores. In general, it is usually appropriate to report as much legitimate and useful information as deemed reasonable, without overinterpreting scores.

Examinees have a right to an accurate, timely, meaningful, and useful report of their test performance. Score reports must be written in language that is understandable to recipients and all appropriate cautions and caveats about misuse of test scores must be clearly and unequivocally stated. All anticipated misuses of the test scores should be clearly labeled; those who provide score reports have an obligation to actively discourage misuse of test scores.

The score scale used to report test results varies by the type of test, the purpose of the examination, and the sophistication of the examinees. For score reporting, the choice of raw scores, percent-correct scores, scaled scores, equated scaled scores, or other types of derived scores should be determined solely on the basis of maximizing communication with the examinee. Whatever score scale is used for reporting, the reported metric should be clearly defined and described in language that is easily understood by the examinee and maximizes the probability of avoiding score misinterpretation. If a passing score is applied to the test results, it is appropriate to generally describe the method or procedures used to establish the passing score and to express the passing score on the same scale as the reported score. If weighted composite scores are reported, the method of establishing the passing score for the composite should be clearly described; likewise, if multiple passing "hurdles" are required, the score report must make this clear. If subscale scores are reported to provide feedback to examinees on their relative strengths and limitations, the subscales must be composed of a sufficient number of test items to ensure a reasonable reliability of the score (i.e., at minimum fifteen to twenty items) and some indication of the standard error of measurement of the subscales should be presented. If subscale scores are reported to examinees solely as feedback on performance and are not used to make pass–fail decisions, the score report must clearly state this fact in language that examinees can easily understand.

The reporting of test scores to examinees is an extremely important step for nearly all types of test development projects. A clear rationale for the type of score report and the reported score scale is essential. Accuracy and absolute clarity of the reported score interpretation are important, as is the active discouragement of score misuse or misinterpretation. Documentation of score reporting activities and their rationale is an important aspect of validity; the score report summarizes, in many important ways, the entire test development program, especially for the examinee and any other legitimate users of test scores.

STEP 11: ITEM BANKING

Secure storage of effective test items is an important step for all on-going testing programs. The process of securely storing test items for potential future use is typically referred to as *item banking* (see Vale, chap. 11, this volume, for a complete discussion of item banking issues).

Because effective test questions are so difficult to develop and so many resources must be used to produce useful test items that perform well, it is sensible to store such questions, together with all their relevant performance data, to reuse such questions on some future form of the examination. Item banks can be as simple as a secure file cabinet using paper copies of examination questions, with the appropriate identifying information and test usage data (such as item difficulty, item discrimination). In practice, most item banking is carried out

using computer software systems, from fairly simple and inexpensive to very complex and expensive software systems.

All item banking systems, if they are to be effective, must be sufficiently flexible and adaptable to serve the needs of test developers. All item banking systems must have the capability to securely store and retrieve test items (and visual materials associated with test items), using all relevant variables useful for the test developer. The required sophistication of the item banking system depends greatly on the type and purpose of the testing program. But all item banking systems must, at minimum, permit the storage, sorting, and retrieval of several variables, such as a unique item identification number, content classification of the test questions (with several subclassifications of content), a cognitive-level classification of the test item, and historical item usage information such as the test form identification (years, dates) of prior use, the item difficulty and item discrimination indices for each prior use (Item Response Theory parameters, if appropriate), and any indication of other items in the bank that should not be used on the same form as a given item. Performance examination test materials and prompts may require more versatile and sophisticated item banking systems than those used for selected-response item formats.

For complex computer-based tests, sophisticated item banking software is most likely required. Some commercial item banking software systems can also serve as test item presentation software for computer-based tests delivered either via the Internet or on secure networks. Developers of complex computer-based tests, especially certain types of adaptive computer-based tests, require very sophisticated software storage and retrieval systems capable of sorting multiple variables simultaneously to build test forms, each of which is representative of a complex test blueprint.

Security of item banks is paramount, no matter what methods are used to store test items and prompts. There are obvious validity issues associated with the security of items stored for potential reuse, but no standards (AERA, APA, NCME, 1999) appear to bear directly on "item banking." If item banks are compromised, the score inferences from such items can also be compromised and the validity evidence for the examination can be decreased seriously. Given the size of the item bank and the stakes associated with the examinations constructed from such item banks, the costs associated with a complete item bank security breach could be extensive. (Some high-stakes testing programs value test items at well over $1,000 per item; almost all test items would be valued at least $300 per item; see Vale, chap. 11, this volume).

Item banking is an important and useful discrete step in test development. After effective test questions are written, edited, reviewed, pretested, and administered to examinees, the items with the most effective item characteristics and the best content should always be preserved for potential reuse on another version or form of the test. It is far too difficult and costly to create effective test items that are used only once. Secure item banking provides a mechanism for convenient, efficient storage and retrieval of test items and may assist test developers in increasing the validity evidence for examinations by helping to control many relevant variables associated with test items.

STEP 12: TEST TECHNICAL REPORT

Every testing program with meaningful consequences for examinees should be systematically documented and summarized in a technical report describing all important aspects of test development, administration, scoring, reporting, and test analyses and evaluation. The technical report is the culminating test development activity and serves the major, but often ignored, purpose of providing thorough documentation of all the validity evidence for a test,

identifies potential threats to validity, and makes recommendations for improvement in the testing program that may strengthen validity evidence.

The *Standards* (AERA, APA, NCME, 1999) address the need for technical reports (chap. 6), emphasizing the documentation aspects of these reports of test development activities (Becker and Pomplun, chap. 30 this volume, address technical reports in detail). Haladyna (2002) presents a validity-based argument for technical reports in relation to the *Standards*.

Test developers are often reluctant to fully document their testing programs. However, the time and effort spent in such documentation activities are rewarded by the ease of retrieval of important validity data and by their support of research efforts on tests (see Haladyna, chap. 32, this volume, for a complete discussion of validity studies). Furthermore, technical reports preserve essential validity evidence for the historical record, including any recommendations for future testing program improvement (Downing & Haladyna, 1997). The overall quality of tests can be improved by focusing careful attention on technical reporting. The examination technical report is also useful in independent evaluations of testing programs, providing a convenient and systematic summary of all important test development activities for review (see Buckendahl & Plake, chap. 31, this volume, for a detailed discussion of the evaluation of testing programs).

One potentially useful model for a technical report is to use the twelve-steps of test development described in this chapter as the major outline headings to organize the report. Depending on the stakes associated with the testing program, each of these twelve-steps requires a significant amount of detailed documentation. Too much documentation is impossible; too little documentation is all too common. The level of detail in the technical report must allow the reader to form clear judgments about the adequacy of each step in test development and about the validity evidence presented for each stage. Some steps may require more documentation than others. For example, for a new high-stakes credentialing examination, it is extremely important to fully document the content definition methods and procedures (and their results), the procedures used to create test specifications, and the methods used to select and train item writers. The methods used to establish the cut score, together with the passing rates associated with implementation of the cut scores, are also important to thoroughly document.

Technical reports must be developed such that all important validity evidence for the testing program is systematically documented in a manner that is easily accessible to all who have a legitimate need to access this information. Reference to the relevant *Standards* (AERA, APA, NCME, 1999) is appropriate in technical documentation, together with the test developer's evaluation of how the test fulfills the recommendations of the *Standards*.

SUMMARY AND CONCLUSION

These twelve-steps for effective test development provide a structured, systematic process for creating effective testing programs of all types. Most of these steps are required for every test. The higher the stakes associated with test scores, the greater the concern for validity (Linn, chap. 2, this volume). Attention to quality control and test security is a pervasive theme running through each of these test development steps. Test development consists of a series of interrelated activities, many of which depend on some prior step or steps of test development. Careful planning and compulsive execution of this detailed plan leads to tests that more validly measure examinee ability or achievement in the well-defined content domain of interest. Adherence to this plan provides validity evidence from multiple sources, as Messick (1989) suggested.

High-quality test development demands great attention to detail. Test validity evidence is increased or decreased, sometimes markedly, as the attention to detail increases or decreases. From the proofreading of item or performance prompt text to the absolute accuracy of test

scoring and reporting, effective quality control methods and procedures must be utilized to ensure that the intended inferences from the test scores are achieved and that CIV is minimized. Systematically following these twelve steps for effective test development helps to ensure maximum test validity evidence for the tests we develop.

ACKNOWLEDGMENTS

The author gratefully acknowledges the research assistant contributions of Cherdsak Iramaneerat, MD, MHPE, to this chapter. Also, the critical reviews and suggestions of Thomas M. Haladyna, PhD, and the insight, inspiration, and thoughtful criticism provided by my students in a graduate-level test development course are appreciated.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington: American Council on Education.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*(2), 109–128.

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). Quality assurance methods for performance-based assessments. *Advances in Health Sciences Education, 8*, 27–47.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M. (2002a). Assessment of knowledge with written test forms. In Norman, G. R., Van der Vleuten, C. P. M., Newble, D. I. (Eds.), *International handbook for research in medical education* (pp. 647–672). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Downing, S. M. (2002b). Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make any difference? *Academic Medicine, 77*(10), S103–104.

Downing, S. M. (2004, April). The effects of violating standard item-writing principles: The impact of flawed test items on classroom achievement tests and students. Paper presentated at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61–82.

Downing, S. M., Lieska, N. G., & Raible, M. D. (2003). Establishing passing standards for classroom achievement tests in medical education: A comparative study of four methods. *Academic Medicine, 78*, S85–87.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220.

Haladyna, T. M. (2002). Supporting documentation: Assuring more valid test score interpretations and uses. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment for all students: Validity, technical adequacy, and implementation* (pp. 89–108). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd Ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 37–50.

Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 51–78.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing*, (pp. 109–127). San Francisco, CA: Jossey-Bass.

Impara J. C., & Plake B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353–366.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77*, 156–161.

Kilminster, S., & Roberts, T. (2003). Standard setting for OSCEs: Trial of borderline approach. *Advances in Health Sciences Education, 8*, 1–9.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*, (2nd Ed.) New York: Springer-Verlag.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. New York: Harcourt Brace.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York: American Council on Education and MacMillan.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.

Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10*, 39–59.

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14*(4), 369–415.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163–184.

Schmeiser, C. B., & Welch, C. (In press). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). New York: American Council on Education and Greenwood.

Spray, J., & Huang, C-Y. (2000). Obtaining test blueprint weights from job analysis surveys. *Journal of Educational Measurement, 37*(3), 187–201.

Thissen, D., & Wainer, H. (2001). An overview of *Test Scoring*. In Thissen, D., & Wainer, H. (Eds.), *Test scoring* (pp. 1–19). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). New York: Springer-Verlag.

Wilkinson, T., Newble, D., & Frampton, D. (2001). Standard setting in an objective structured clinical examination: Use of global ratings of borderline performance to determine the passing score. *Medical Education, 35*(11), 1043–1049.

Zieky, M. J. (1997). Is the Angoff method really fundamentally flawed? *CLEAR Exam Review, 7*(2), 30–33.

Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980's. In G. J. Cizek, (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum Associates.